ChinaXiv合作期刊
图书情报工作
第63卷 第8期 2019年4月

中文学术期刊论文的引文模式研究*
——以2006-2008年图书情报领域期刊论文为例

熊泽泉[1,2] 段宇锋[1]
[1] 华东师范大学经济与管理学部 上海 200241 [2] 华东师范大学图书馆 上海 200241

摘要：[目的/意义]探索中文学术期刊论文的引文模式及时间窗口的选择对引文模式的影响，建立引文模式的分析框架。[方法/过程]以2006-2008年出版的图书情报领域期刊论文作为研究对象，采用两步聚类法对单篇论文在7年内的绝对被引量与相对被引量进行聚类分析，研究论文主要特征因子与引文模式的相关性。[结果/结论]在绝对被引量视角下，期刊论文均表现为先上升后下降的经典引文模式；在相对下载量视角下，期刊论文共有6种引文模式，其中3种可以归纳为经典引文模式，另外3种分别为"类睡美人型"、正偏型和马拉松型。相对被引量视角下，首年被引量与总被引量呈现了中等甚至较强的相关性，并且平均被引量越高，相关性越强，绝对被引量视角下的结果正好相反。结果表明，期刊论文的初始被引量与总被引量的相关性高低主要取决于引文曲线的峰度而非总被引量的大小。

关键词：被引量 引文模式 时间窗口 老化 相关性
分类号：G25
DOI：10.13266/j.issn.0252-3116.2019.08.013

1 前言

长期以来，被引次数在论文学术影响力评价中起着举足轻重的作用。这一评价指标关注的是论文静态的总被引次数，未考虑其动态变化，是一种只看"结果"不重"过程"的评价方法。然而，总被引次数一样的论文，其引用的过程和机理可能存在差异，两个极端的例子是"昙花一现"型论文[1-2]和"睡美人"型论文[2-3]。这两种类型的论文不仅具有不同的学术价值，其随后数年所获得的被引次数也可能存在较大差异。因此，研究者不仅需要关注论文的"结果"，同样也需对产生"结果"的"过程"及机理进行研究，才能对论文的学术影响力正确地进行评价。

论文引用的"过程"一般可以采用被引次数随时间变化的曲线来进行描述。这种动态反映被引次数随时间变化的曲线称为引文曲线（citation curve）[4]或引文模式（citation pattern）[5]，也有学者称其为引文生命周期（citation life cycle）[6]、引文轨迹[7]或引文历史[8]。引文模式从引用的时间分布入手，分析引用的历史过程，以此反映论文的影响力[9]。根据所采用方法的不同，对于引文模式的研究大致可以分为如下两类：

（1）采用曲线拟合的方法建立引文模式。譬如A. Avramescu通过拟合曲线，发现单篇论文可能具有3种引文模式，即常态型、指数增长型、昙花一现型；并且进一步归纳出引文曲线的一般公式：$C(t) = C_0[\exp(-\alpha t) - \exp(-m\alpha t)]$，$m>1$，其中$C_0$表示振幅，$\alpha$是时间常量，m为初始增量；同时，A. Avramescu也承认，该公式无法描述所有的引文曲线类型，比如双峰型曲线，或者其他不规则曲线[10]。李江等以341位诺贝尔奖获得者作为研究对象，通过拟合每一获奖者自发表第一篇论文起截止到2011年的引文曲线，构建了引文曲线分析框架，该框架中包含两大类引文曲线，一类是规则引文曲线（包括经典引文曲线和指数增长引文曲线），另一类为不规则引文曲线（包括睡美人引文曲线、双峰引文曲线和波型引文曲线）[9]。不过，作者认为该分析框架仅适用于有一定年限、有一定影响的学者和论文。

（2）另外的一些研究采用聚类的方法分析论文的

* 本文系ISTIC-ELSEVIER期刊评价研究中心开放基金资助项目"CiteScore与JCR影响因子的学科差异性研究"研究成果之一。
作者简介：熊泽泉（ORCID：0000-0002-4349-371X），馆员，博士研究生；段宇锋（ORCID：0000-0002-4349-371X），教授，博士生导师，通讯作者，E-mail：yfduan@infor.ecnu.edu.cn。
收稿日期：2018-08-08 修回日期：2018-11-16 本文起止页码：107-115 本文责任编辑：易飞

引文模式。譬如,E. S. Aversa 采用 K 均值聚类法,对 1972 年出版的 400 篇高被引论文在出版后 8 年内的引文曲线进行聚类分析,发现在高被引论文中存在两种引文模式,即"延迟增长——缓慢下降"型和"立即增长——快速下降"型,前者获得的总被引量更高[5],并且认为这一结果验证了 D. D. S. Price 提出的"被引量较低的论文老化速率更快"这一结论[11];V. Cano 和 N. C. Lind 对 10 篇高被引文献和 10 篇中低被引文献在发表后 25 年内的引文生命周期进行研究,也发现了两种引文模式,其中模式 A 由高被引论文和中低被引论文混合组成,在第六年其被引量的增长速率显著下降,累计被引量曲线呈双(折)线型,模式 B 均由高被引论文组成,其被引量具有稳定的增长速率,累计被引曲线呈单线型[6];S. E. Baumgartner 和 L. Leydesdorff 采用"组基轨迹建模" Group-based trajectory modeling (GBTM)方法对 1996 年出版的 JASIS、JACS、Cell、Gene、Science、Nature 等 6 本期刊以及病毒学的 24 本期刊中的论文,在 15 年内的被引曲线进行聚类分析,同样发现两类引文模式,并将其归纳为出版后数年达到被引高峰,然后逐渐老化的"短暂知识要求(transient knowledge claims)"和出版后十年内持续被引的"粘滞知识要求(sticky knowledge claim)"[12]。不过,这些聚类方法都必须事先对聚类数进行人为指定,然后再根据聚类结果的优劣选择最佳的聚类数。

上述研究中,有的选择特定的数据作为研究对象,比如高被引论文、诺贝尔奖获得者论文,有的采用了不同的数据处理方法或不同的时间窗口,因此得出的结论并不一致。同时,已有研究主要以可被广泛使用的英文文献作为研究对象,尚没有学者关注中文学术期刊论文的引文模式,是否中文学术期刊论文会因使用人群的局限而存在不同的引文模式?本研究将通过对图书情报领域中文学术期刊论文引文模式的分析,来探究中文期刊论文引文模式的特点及影响因素。同时,在笔者先前的研究中,分别对中文期刊论文的下载模式[13]、被引量与下载量的动态相关性[14]进行了分析,本文采用相同的数据集对期刊论文的引文模式进行研究,与上述系列研究构成了一个完整的研究体系,并为后续被引量预测模型的研究奠定基础。

2 数据和方法

2.1 数据来源与处理

以中国最大的期刊数据库——中国学术期刊(网络版)全文数据库作为数据源,选择其中的《大学图书馆学报》《情报科学》《情报理论与实践》等 11 种图书情报领域期刊在 2006-2008 年发表,且在 2015 年 12 月 31 日前获得过被引的 9 066 篇论文作为研究对象。期刊选择依据主要是创刊时间较长、在数据库中收录完整,且其出版时间和上线时间基本一致,从而能够获得较为真实的下载量及被引量数据,便于后续研究。《图书情报工作》《中国图书馆学报》等期刊因为出版到上线的滞后期较长,未选择其作为研究对象。将该原始数据集命名为 DataSet1。由于中国学术期刊(网络版)全文数据库中单篇论文下载量的数据从 2006 年才开始统计,为了与后续研究被引量与下载量相关性时数据的时间切面保持一致,同时已有研究证明 7-8 年的时间窗口能够基本反映论文的引文模式[5],因此本研究统一采用从 2006 年 1 月 1 日至 2015 年 12 月 31 日单篇论文的被引量数据。

DataSet1 中,每篇论文所涉及的数据包含论文的基本题录信息以及该论文在 2006-2015 年每一自然年的被引量,加总每一自然年的被引量,得到每篇论文自出版时到 2015 年 12 月 31 日的总被引量;由于不同论文出版月份不同,有的在年初出版,有的在年末出版,因此出版月份较晚的论文在出版当年的被引量无法体现其真实的年均被引量,为了更加准确地呈现论文在出版后 1 年内的被引量,本文假设每篇论文被引量在一年的不同月份不存在差异,采用如下公式来进行计算:$C'_{Y+1} = C_Y + \frac{C_{Y+1}}{12} \times (12 - M)$,出版后第 2 年的被引量为 $C'_{Y+2} = \frac{C_{Y+1}}{12} \times M + \frac{C_{Y+2}}{12} \times (12 - M)$,其中 M 表示论文出版月份,$C_Y$ 表示论文在 第 Y 年的被引量,C_{Y+1} 表示论文在发表 Y+1 年的被引量, C'_{Y+1} 表示计算获得的论文在发表 1 年后的实际被引量。采用同样方法获得每篇论文出版后 3-7 年内的被引量,作为新数据集 DataSet2(2008 年发表的论文截至 2015 年 12 月 31 日只有 7 年的被引数据,因此将所有论文统计年限统一为 7 年);同时,对 7 年的被引量进行归一化处理,计算每篇论文每年被引量占其 7 年总被引量的百分比,作为归一化数据集 DataSet3。通过上述数据处理,本文得到了 3 个不同的数据集——原始数据集 DataSet1、绝对被引数据集 DataSet2、归一化数据集 DataSet3。以 2008 年 5 月发表于《情报科学》的论文《自然语言检索中的中文分词技术研究进展及应用》为例,在 DataSet2 中,$C'_{2008+1}{}' = C_{2008} + \frac{C_{2008+1}}{12} \times (12-7)$

$=3+\frac{10}{12}\times5=7.17$，$C'_{2008+2}=\frac{C_{2008+1}}{12}\times7+\frac{C_{2008+2}}{12}\times(12-7)=\frac{10}{12}\times7+\frac{14}{12}\times5=11.67$。相应地，在 DataSet3 中，$R_{2008+1}=7.17/(7.17+11.67+12.33+11.25+11.75+10.00)+10.42)\times100\%=9.61\%$，$R_{2008+2}=11.67/(69.75+33.75+26.25+21.25+12.50+20.50)+16.50)\times100\%=15.64\%$，其被引量在3个数据集中的不同表示方式如表1所示：

表1 论文被引量在不同数据集中的表示方式的示例

DataSet 1							
2008(年)	2009	2010	2011	2012	2013	2014	2015
3	10	14	10	13	10	10	11
DataSet 2							
第1年	第2年	第3年	第4年	第5年	第6年	第7年	
7.17	11.67	12.33	11.25	11.75	10.00	10.42	
DataSet 3							
第1年	第2年	第3年	第4年	第5年	第6年	第7年	
9.61%	15.64%	16.54%	15.08%	15.75%	13.41%	13.97%	

2.2 分析方法

(1)聚类分析。采用IBM SPSS Statistics 23提供的两步聚类法(two-step cluster)，分别从论文的绝对被引量(即论文每年的实际被引频次，采用绝对被引数据集Dataset 2进行)和相对被引量(即论文每年实际被引频次占总被引频次的百分比，采用归一化数据集Dataset 3进行分析)两个角度，对Dataset2和Dataset3中的论文进行聚类分析，探索不同的引文模式。两步聚类法是一种新型的分层聚类算法，可处理大样本的连续变量和分类变量，因其聚类步骤包含预聚类和子簇聚类两步而得名，并且该聚类方法可自动确定聚类数目。具体聚类步骤为：分别选择DataSet2(或DataSet3)中每年的被引次数作为连续变量，聚类准则采用施瓦兹贝叶斯准则(BIC)，由于之前对数据已经进行了清理，因此对离群值不再使用噪声处理，评估字段采用唯一的文件识别号，并勾选创建聚类成员变量，最终得到每一篇论文所属类群。

(2)相关性分析。采用Spearman相关系数对不同的引文模式中论文被引量与论文标题长度、作者数量、关键词数量、期刊复合影响因子以及初始被引量等进行相关性分析。

3 研究结果

3.1 基于论文绝对被引量的引文模式

采用两步聚类法对DataSet2中每年的绝对被引量进行聚类分析，发现样本可以聚类为3种引文模式，分别命名为模式1、模式2和模式3(见图1)，从这3种引文模式可以得到如下两点发现：

(1)模式1、模式2和模式3的变化趋势基本一致，均呈先上升后下降的模式，但总被引量相对高的论文(模式1)达到其被引峰值较晚(3年)，模式2和模式3更早地达到了其被引峰值。这一发现表明，论文老化现象在学术期刊论文中普遍存在，高被引论文进入老化期较晚[5, 11－12]。

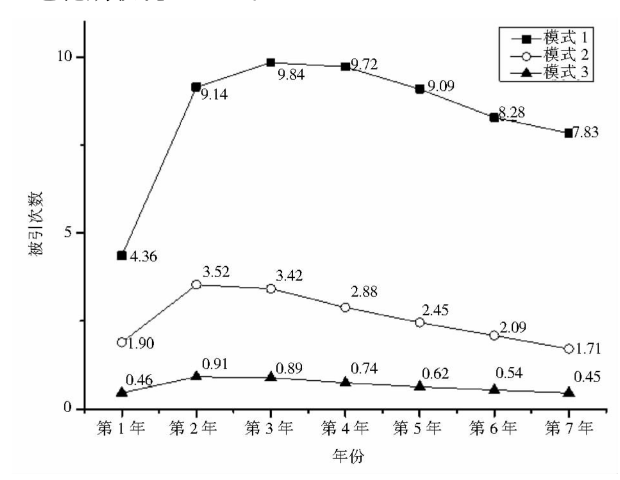


图1 三种绝对被引量模式逐年变化趋势

(2)计算得到每一模式的样本量、总被引量及篇均被引量，结果如表2所示。可以看出，模式1虽然只占到样本总量的4%左右，但其总被引量占比达到了22%，篇均总被引67.52次；而占样本总量64.31%的模式3，论文篇均总被引只有5.15次，大部分论文年均被引不到1次，这一结果与D. W. Aksnes等的结果一致[15－16]。

表2 DataSet2中三种引文模式的样本量、总被引量及篇均被引量

类别	样本数(篇)	样本比例	篇均被引量(次)	总被引量(次)	总被引量占比
模式1	366	4.04%	67.52	24 711	22.11%
模式2	2 870	31.66%	19.87	57 031	51.02%
模式3	5 830	64.31%	5.15	30 048	26.88%

3.2 基于相对被引量的引文模式

采用相同方法对DataSet3中的数据进行聚类分析。有68篇论文由于只在2015年被引用，经数据转换后，其7年的被引量均为0，无法进行相对被引量计算，因此不能归入任何模式。剩余的8 998个样本共聚为如图2所示的6种引文模式，6种引文模式的每年被引量占比及被引量均值及样本数、总被引量及篇均被引量分别见表3和表4。

模式 A:被引量快速上升后迅速下降,然后又缓慢上升。该模式论文数量为 858 篇,占总样本数的 9.46%。

模式 B:被引量缓慢上升后再缓慢下降;该模式论文数量为 2 517 篇,占总样本数的 27.76%。

模式 C:变化趋势与模式 A 类似,即快速上升后迅速下降,然后又缓慢上升,但是峰值比模式 A 滞后;该模式论文数量为 1 087 篇,占总样本数的 11.99%。

模式 D:缓慢上升后再缓慢下降,变化趋势与模式 B 类似,但变化更为平缓,几乎无峰出现;该模式论文数量为 3 456 篇,占总样本数的 38.12%。

模式 E:变化趋势与模式 A 和模式 C 类似,但峰值比模式 C 更加滞后;该模式论文数量为 639 篇,占总样本数的 7.05%。

模式 F:出版后的最初几年几乎无被引,后被引量突然增加。该模式样本数量为 441 篇,占样本总数的 4.86%。

从相对被引量的变化趋势来看,模式 A、模式 C、模式 E 可以近似看作是同一条变化曲线沿着 X 轴的平移,曲线趋势基本一致,仅峰值出现时间有所不同。

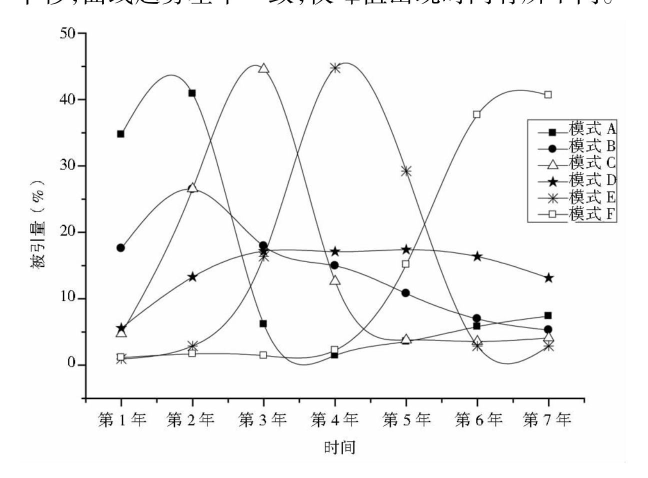


图 2 归一化数据集 6 种引文模式变化趋势

表 3 归一化数据集中 6 种引文模式每年被引量占比及被引量均值

类别		第一年	第二年	第三年	第四年	第五年	第六年	第七年
模式 A	相对被引量	34.73%	40.86%	6.16%	1.48%	3.57%	5.803	7.40%
	绝对被引量	1.08	1.18	0.24	0.09	0.14	0.22	0.30
模式 B	相对被引量	17.57%	26.48%	17.95%	14.96%	10.80%	6.96%	5.30%
	绝对被引量	1.98	3.26	2.38	1.88	1.38	0.98	0.77
模式 C	相对被引量	4.75%	26.58%	44.56%	12.62%	3.79%	3.57%	4.14%
	绝对被引量	0.37	1.65	2.35	0.88	0.34	0.30	0.31
模式 D	相对被引量	5.59%	13.28%	17.16%	17.09%	17.39%	16.36%	13.13%
	绝对被引量	0.97	2.21	2.73	2.75	2.72	2.51	2.15
模式 E	相对被引量	0.96%	2.89%	16.38%	44.76%	29.23%	2.88%	2.91%
	绝对被引量	0.06	0.16	0.56	1.29	0.74	0.11	0.12
模式 F	相对被引量	1.15%	1.70%	1.44%	2.26%	15.16%	37.66%	40.64%
	绝对被引量	0.06	0.09	0.09	0.12	0.40	0.91	0.84

表 4 归一化数据集中 6 种引文模式的样本数、总被引量及篇均被引量

类别	样本数(篇)	样本比例	总被引量(次)	总被引量占比	篇均被引量(次)
模式 A	858	9.54%	3 025	2.71%	3.53
模式 B	2 517	27.97%	34 350	30.74%	13.65
模式 C	1 087	12.08%	7 309	6.54%	6.72
模式 D	3 456	38.41%	63 492	57.22%	18.50
模式 E	639	7.10%	2 163	1.94%	3.38
模式 F	441	4.90%	1 401	1.25%	3.18

3.3 绝对和相对引文模式的矩阵分析

上文基于绝对被引量得到了 3 种引文模式,基于相对被引量得到了 6 种引文模式,那么属于不同绝对引文模式的论文,其相对引文模式又如何呢?本文对绝对引文模式和相对引文模式进行了矩阵分析(见表 5)。从表 5 可以看出,代表了高被引论文的模式 1 中,其相对引文模式主要表现为模式 B 和模式 D,而模式 A、模式 C 和模式 F 各只有 1 篇论文,分别为 2006 年出版的“我国图书馆精神研究综述”(被引次数为 14、9.5、3.5、1.5、1、1、1)、2008 年出版的“国外图书馆的信息共享空间”(被引次数为 3.9、11.3、13.8、11、0.2、2.2、3.8)、2008 年出版的“文献数据库中书目信息共现挖掘系统的开发”(被引次数为 0.7、3.7、8.3、8.7、12.7、19.7、38)。

研究者普遍将前 1% 的论文认为是高被引论文,参照这一标准,本文将论文按总被引量排序,共获得 90 篇高被引论文。这 90 篇论文共获得 10 582 次被引,占总被引量的 9.47%。它们均属于前述绝对引文

模式中的模式 1,而从相对引文模式来看,11 篇属于模式 B,78 篇属于模式 D,只有 1 篇属于模式 F,该篇论文为上文提到的“文献数据库中书目信息共现挖掘系统的开发”。

通过上述数据分析,本文认为:模式 B 和模式 D 是论文获得高被引的必要条件,但并非充分条件。

表 5 两种数据集不同下载模式的样本量矩阵

	模式 1	模式 2	模式 3	总计
模式 A	1(0.01%)	66(0.73%)	791(8.79%)	858(9.54%)
模式 B	95(1.06%)	1 216(13.51%)	1 206(13.40%)	2 517(27.97%)
模式 C	1(0.01%)	179(1.99%)	907(10.08%)	1 087(12.08%)
模式 D	268(2.98%)	1 395(15.50%)	1 793(19.93%)	3 456(38.41%)
模式 E	0(0.00%)	11(0.12%)	628(6.98%)	639(7.10%)
模式 F	1(0.01%)	3(0.03%)	437(4.86%)	441(4.90%)
总计	366(4.07%)	2 870(31.90%)	5 762(64.04%)	8 998(100.00%)

为了进一步研究高被引论文的特征,采用 Spearman 相关系数对高被引论文的总被引量与题名长度、作者数量、关键词数量、复合影响因子等指标进行相关性分析(见表 6),发现高被引论文的总被引量仅与作者数量存在弱相关性。而首年被引量与总被引量不存在相关性,直到第三年的累计被引量才与总被引量呈弱相关性。

表 6 高被引论文总被引量与主要论文特征因子的相关系数

	题名长度	作者数量	关键词数量	复合影响因子	首年被引量	二年累计被引量	三年累计被引量
相关系数	-0.062	0.304 **	-0.029	0.100	0.085	0.243 *	0.371 **
显著性	0.560	0.004	0.788	0.350	0.427	0.021	0.000

注:**相关性在 0.01 层上显著(双尾)

3.4 不同引文模式下论文被引量与论文特征的相关性分析

一般认为,论文被引量的高低直接反映了被引论文质量的好坏[6],一些研究显示,论文的特征与其被引量存在一定的关系,但这类关系存在学科、语种上的差异[17]。因此,本部分研究采用 Spearman 相关系数分析不同引文模式下,论文总被引量与论文特征之间的相关性,同时为了研究先期被引量对总被引量的影响,本文也分析了论文首年被引量、二年累计被引量、三年累计被引量与总被引量的相关性(见表 7 和表 8)。

结果表明:①各引文模式中的论文,被引量与题名长度、作者数量、关键词数量、复合影响因子等特征均不存在显著的相关性或相关性非常弱;②在相对被引量视角下,首年被引量与总被引量呈现了中等甚至较强的相关性,并且平均被引量越高,相关性越强。

表 7 不同(绝对)引文模式下论文被引量与论文特征的相关系数

	模式 1	模式 2	模式 3	样本总体
题名长度	-0.057	-0.044 *	-0.020	-0.037 **
作者数量	0.132 *	0.039 *	0.097 **	0.090 **
关键词数量	-0.005	-0.038 *	0.102 **	0.064 **
复合影响因子	0.151 **	0.092 **	0.061 **	0.065 **
首年被引量	0.106 *	0.082 **	0.344 **	0.632 **
二年累计被引量	0.290 *	0.363 **	0.576 **	0.805 **
三年累计被引量	0.454 **	0.587 **	0.730 **	0.889 **

注:**. 相关性在 0.01 层上显著(双尾)

*. 相关性在 0.05 层上显著(双尾)

表 8 不同(相对)引文模式下论文被引量与论文特征的相关系数

	模式 A	模式 B	模式 C	模式 D	模式 E	模式 F	样本总体
题名长度	-0.038	-0.054	-0.012	-0.020	-0.028	0.035	-0.037 **
作者数量	0.000	-0.016	0.055	0.100 **	0.065	0.101 *	0.090 **
关键词数量	0.049	0.014	0.031	0.012	0.126 **	0.126 **	0.064 **
复合影响因子	-0.102 **	0.007	0.041	0.067 **	0.058	0.109 **	0.065 **
首年被引量	0.594 **	0.726 **	0.595 **	0.690 **	0.397 **	0.411 **	0.632 **
二年累计被引量	0.833 **	0.932 **	0.843 **	0.859 **	0.625 **	0.502 **	0.805 **
三年累计被引量	0.847 **	0.959 **	0.953 **	0.920 **	0.737 **	0.614 **	0.889 **

注:**. 相关性在 0.01 层上显著(双尾)

*. 相关性在 0.05 层上显著(双尾)

3.5 主题分析

为了研究论文主题对引文模式的影响,本文将每一论文的主题分类号取至第三级,选择论文数量超过 100 篇的三级主题分类号进行统计,计算每一主题下的论文数量、该主题下模式 1 的论文数量、该主题下模式 1 论文占该主题论文数量的比例,以及该主题下模式 1 论文占模式 1 总论文数的比例(见表 9)。

结果显示,论文的主题与其绝对被引模式有一定

的相关性。信息理论(G201)中模式 1 的论文比例远远高于其他主题,达到 15.74%;图书馆学(G250)虽然在模式 1 中占比远高于其他主题(18.31%),但主要是由于该主题论文总数量多,而该主题内模式 1 的论文占比仅为 4.48%;各种文献工作(G255)、文献学(G256)和信息产业经济(F49)这 3 个主题不存在模式 1 论文。

同时,本文发现各主题论文数量与模式 1 论文数量呈显著相关性(Spearman 相关系数为 0.631),这一结果暗示,论文数量越高的主题,越有可能出现高被引论文。

表 9 不同主题下模式 1 论文数量及占比情况

分类号	主题	论文数量(篇)	模式 1 论文数量(篇)	论文占比(相对主题)	论文占比(相对模式 1)
G250	图书馆学	1 495	67	4.48%	18.31%
G252	读者工作	898	44	4.90%	12.02%
G258	各类型图书馆	719	42	5.84%	11.48%
G251	图书馆管理	556	21	3.78%	5.74%
TP39	计算机的应用	491	22	4.48%	6.01%
G354	情报检索	307	5	1.63%	1.37%
TP31	计算机软件	302	11	3.64%	3.01%
G253	藏书建设和藏书组织	298	6	2.01%	1.64%
G259	世界各国图书馆事业	280	7	2.50%	1.91%
G254	文献标引与编目	272	7	2.57%	1.91%
G350	情报学	254	12	4.72%	3.28%
F270	企业经济理论和方法	242	12	4.96%	3.28%
F272	企业计划与经营决策	238	5	2.10%	1.37%
G203	信息资源及其管理	215	8	3.72%	2.19%
G353	情报资料的处理	204	20	9.80%	5.46%
G255	各种文献工作	181	0	0.00%	0.00%
G256	文献学	181	0	0.00%	0.00%
F49	信息产业经济(总论)	110	0	0.00%	0.00%
G201	信息理论	108	17	15.74%	4.64%
总论文数	9 066	366	4.04%	100.00%	

4 讨论

4.1 论文引文模式的类型

本次研究分别基于绝对被引量和相对被引量进行了引文模式分析,并对两类引文模式进行了矩阵分析,从而获得了一种研究引文模式的新视角。

基于绝对被引量,本研究获得了 3 种引文模式,这 3 种模式变化趋势基本一致,均表现为先上升后下降的"经典引文模式",只是在绝对数量和变化速率上有所不同,可见在这一视角下,绝对数量的大小在聚类时所获得的权重更大,数据集聚类成高被引论文、低被引论文和中度被引论文这 3 种模式,而数值的变动情况则被掩盖在这 3 种模式中;同时,这 3 种引文模式的共同趋势,也进一步验证了文献老化规律的普遍存在。

本文在相对被引量视角下,获得了 6 种引文模式,其中模式 A、模式 C、模式 E 的曲线峰度与偏度基本一致,且均呈正态分布,可以看成是同一条曲线沿着 X 轴的平移,仅峰值出现的时间有所不同,可以将这 3 种模式归纳为同一类引文模式,即典型的单峰型,表现为被引次数的快速上升和快速下降,这也与 A. Avramescu 提出的常态型引文模式一致[10];而且从矩阵分析的结果看,几乎不存在模式 A1、模式 C1 和模式 E1 的情况,即典型的单峰型引文曲线几乎不存在高被引论文。模式 F 在出版后 4 年基本上没有被引用,从第 5 年开始被引量逐渐上升,且在本研究选择的时间范围内并未显示老化趋势,其曲线形状类似于"睡美人"引文曲线,但是如果从绝对被引量的视角来评价模式 F 的论文时,发现多数论文的总被引次数少于 10 次,参考 A. Raan 提出的"睡美人"定量标准[3]——"沉睡"期年均被引量≤2,"苏醒"后 4 年内总被引量 >20,显然模式 F 中的论文并不属于一般意义上的的"睡美人",因此本文将其定义为"类睡美人"型——相对被引量变化趋势与"睡美人"论文类似,呈现出版后一段时间被引量几乎为 0,后被引量突然迅速增长,但数据积累时间

较短，其“苏醒”后的绝对被引量也无法达到睡美人标准。模式 B 虽然也存在明显的波峰，但其峰度和偏度均与其他引文曲线存在显著差异，峰度小于上述曲线，偏度呈明显的正偏。模式 D 则没有明显的波峰，呈现缓慢上升 – 缓慢下降的变化趋势，年均被引占比最高不超过 20%，这一模式类似于 E. S. Aversa 发现的“延迟增长 – 缓慢下降”型的引文曲线[5]，或者称之为马拉松型[8]；同时本文发现，如果从绝对被引量的角度来看，代表高被引论文的模式 1 中，模式 D 的比例也最高，即被引量高的论文在被引量变化趋势上主要呈现为模式 D，表现为比低被引论文更慢的老化趋势[5, 16, 18 – 19]，对 90 篇高被引论文的分析结果也得到了同样的结果。由此可以推测，对于大部分期刊论文，绝对被引量越高，老化趋势越为缓慢，其相对被引量曲线越平缓。但反之并不成立。

4.2 时间窗口选择对于论文引文模式的影响

引文模式的分类研究，涉及两个重要的影响因素，一个是被引次数，另一个是被引时间窗口。被引时间窗口是指选择论文出版后多少年被的被引量来测定其学术影响力[20]。对于同一篇论文，时间窗口选择不同，所得出的结论也有可能不一致。一个典型的例子就是“睡美人”型论文，如果选择的时间窗口较小，这一类型的论文就有可能被埋没[20]，同样地，当研究者把时间窗口放大时，原来单峰型的引文曲线可能演化成双峰型，原来指数增长型的引文曲线也可能随着被引量的减少而演化成常态的单峰型。除此之外，时间窗口较短也可能导致在不同学科领域之间的不公平对待[21]，一般而言，社会科学类论文的被引都显著滞后于自然科学论文。然而，在实际操作中，研究者往往无法等到数十年后再来对某篇论文或某个学者的影响力进行评价，而不得不利用 3 – 5 年的时间窗口来进行学术影响力测定。这就提出了一个问题，短期时间窗口所得出的结论准确性到底有多高？

在已有的一些相关研究中，由于样本量的不同或者评价标准的不一，得出的结论也存在着差异。J. Adams 分析了英国 8 258 篇不同学科的论文后发现，初始被引量（1 – 2 年）和随后 3 – 10 年的被引量有着显著的相关性，最小的相关系数达 0.653[22]；J. M. Levitt 和 M. Thelwall 在对 1970 年出版的 6 个学科中的 87 篇高被引论文进行研究后发现，有 4 个学科中的早期被引量（6 年）与总被引量的相关系数超过 0.42[18]；J. Wang 以 Web of Science 中 1980 年出版的 358 100 篇论文作为研究对象，分析了论文出版后 1 – 31 年的累积被引量与总被引量的相关性，发现相关系数从第 1 年的 0.266 快速上升到第 3 年的 0.756，然后缓慢上升，因此作者认为如果对样本总体进行评价，相关系数达到 0.8 已经足够的前提下，选择 4 年的时间窗口就已经足够[20]；并且作者发现，对于高被引论文，初始被引量与总被引量的相关系数更低。

在本研究中，高被引论文三年累计被引量与总被引量的相关系数为 0.371，依据绝对被引量聚类得到的模式 1、模式 2 和模式 3 三年累计被引量与总被引量的相关系数分别为 0.454、0.570 和 0.730，即总被引量越高，初始被引量与总被引量的相关性越低，这与 J. Wang 所得出的结论一致[20]；而基于相对被引量聚类而成的 6 种引文模式的三年累计被引量与总被引量的相关性从 0.614 到 0.959 不等，其中模式 B 和模式 D 中的论文相关系数最高，并且首年的被引量与总被引量的相关系数就达到了 0.7 左右，而这两种模式又是高被引论文最多的两种模式。可见，初始被引量与总被引量的相关性高低取决于引文曲线的峰度而非绝对被引量的大小，之前研究得出的高被引论文的初始被引量与总被引量相关系数更低的结论并不完全正确：被引量变化较为缓慢的高被引论文其初始被引量与总被引量会呈现更高的相关性，而属于指数增长型或者是“睡美人”型引文曲线的高被引论文，其初始被引量与总被引量相关性则较低。

4.3 引文模式的分析框架

被引次数是进行引文模式分类的重要依据，可以用绝对被引量和相对被引量两种方式表示，一般而言，研究者通常会选择绝对被引量作为研究对象，获得其随时间的变化规律[6, 12, 23 – 24]；而在涉及到论文老化规律时，研究者一般会以所选时间框架内每年的相对被引量[5, 18, 25 – 26]或累计的相对被引量[20]作为研究对象，进而得出半衰期等指标作为其老化率的评价依据。这两种方式各有优缺点，以绝对被引量作为研究对象时，能够不受时间窗口的限制获得被引量实时的变化规律，而且对于高被引论文和低被引论文有很好的区分度，但正因为如此，无法在总被引量处于相同水平时识别出具有不同相对量变化趋势的论文。比如在模式 1 中，一些论文的被引量实际是在持续上升的，但这一类型的论文却被淹没在大量具有常态引文轨迹的论文中而无法被发现。

而以相对被引量作为研究对象时，在设定的时间窗口内，研究者可以忽略论文绝对被引量大小的差异，而专注于研究其逐年变化趋势，但同时由于下载量被

标准化,这种视角无法对绝对数量进行区分,造成一些高被引论文和低被引论文因变化趋势类似被归为一类,不利于合理地评价。譬如,在李江等的研究中,作者认为以单篇论文为对象时,95%的论文年均被引次数小于10,被引曲线呈不规则的波型或无法绘制引文曲线,大量的波型引文曲线淹没了其他类型的引文曲线,使得统计分析的有效性大大降低,因此作者认为,其提出的引文曲线分析框架适用于有一定影响力的研究对象(单篇论文或单个作者)[9]。这一观点正是从绝对被引次数的视角得出的。但事实上,当研究者选择一个合适的时间窗口,将年被引次数转换为相对值时,被引次数的绝对大小差异便可以忽略,而主要聚焦曲线整体的变化趋势,这样依然可以观测到不同模式的引文曲线。然而,这种方式可能会造成部分高被引论文因为数据被标准化而无法被发现,致使其淹没在大量具有相同变化趋势的中度被引论文和低被引论文中。

本文通过对绝对引文模式与相对引文模式进行矩阵分析,构建了分析论文老化规律的一种新的研究框架,能够快速地对不同的引文模式进行识别,有助于更加全面地分析论文的老化规律。该框架不仅能够识别出高被引论文和低被引论文,同时也对论文获得被引的过程作了区分,是一种既重"结果",也重"过程"的评价方法。在这种新的分析框架内,本文发现,对于大部分期刊论文,绝对被引量越高,其相对被引量曲线越平缓,但反之并不成立;同时,本文发现,期刊论文的初始被引量与总被引量的相关性高低取决于引文曲线的峰度而非总被引量的大小,这些发现将有助于进一步利用学术论文的短期被引量来预测其未来被引量及老化模式。

5 不足与展望

本研究采用图书情报领域期刊论文作为研究对象,构建了一种更加全面的的引文模式分析框架,但对其他领域论文是否适用,是否可以推广到所有学科领域,仍有待进一步研究。

在对期刊论文下载模式[13]、引文模式进行分析的基础上,进一步对下载量和引文量的动态相关性以及不同下载和引文模式下论文下载量和被引量的相关性进行研究[14],有助于利用论文早期下载量与被引量等论文特征因子预测论文长期被引量,从而较早地发现未来可能被高被引的论文或高被引学者,助力于科研管理人员的管理与决策活动。在后续研究中,笔者将以上述系列研究为基础,探索更高效的被引量预测模型。

致谢:本研究数据由中国知网(CNKI)提供。华东师范大学图书馆杨莉老师在本文的数据分析过程中给予了耐心的指导,在此一并表示感谢。

参考文献:

[1] VAN DALEN H P, HENKENS K. Signals in science - on the importance of signaling in gaining attention in science[J]. Scientometrics, 2005, 64(2): 209-233.

[2] 李江. 科学中的"睡美人"与"昙花一现"现象评述[J]. 大学图书馆学报, 2016(3): 38-43.

[3] VAN RAAN A. Sleeping beauties in science[J]. Scientometrics, 2004, 59(3): 467-472.

[4] GARFIELD E. More delayed recognition. 1. examples from the genetics of color-blindness, the entropy of short-term-memory, phosphoinositides, and polymer [J]. Current contents, 1989, 38: 3-8.

[5] AVERSA E S. Citation patterns of highly cited papers and their relationship to literature aging - a study of the working literature[J]. Scientometrics, 1985, 7(3/6): 383-389.

[6] CANO V, LIND N C. Citation life-cycles of 10 citation-classics [J]. Scientometrics, 1991, 22(2): 297-312.

[7] 杜建,武夷山. 文献引文轨迹:分类及测度[J]. 情报理论与实践, 2015(7): 52-58.

[8] COLAVIZZA G, FRANCESCHET M. Clustering citation histories in the physical review[J]. Journal of informetrics, 2016, 10(4): 1037-1051.

[9] 李江,姜明利,李玥婷. 引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例[J]. 中国图书馆学报, 2014(2): 41-49.

[10] AVRAMESCU A. Actuality and obsolescence of scientific literature [J]. Journal of the American Society for Information Science, 1979, 30(5): 296-303.

[11] PRICE D D S. A general theory of bibliometric and other cumulative advantage processes[J]. Journal of the American Society for Information Science, 1976, 27(5): 292-306.

[12] BAUMGARTNER S E, LEYDESDORFF L. Group-based trajectory modeling (GBTM) of citations in scholarly literature: dynamic qualities of "transient" and "sticky knowledge claims"[J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 797-811.

[13] DUAN Y, XIONG Z. Download patterns of journal papers and their influencing factors[J]. Scientometrics, 2017, 112(3): 1761-1775.

[14] 熊泽泉,段宇锋. 论文早期下载量可否预测后期被引量?——以图书情报领域期刊为例[J]. 图书情报知识, 2018(4): 32-42.

[15] AKSNES D W, SIVERTSEN G. The effect of highly cited papers

on national citation indicators[J]. Scientometrics, 2004, 59(2): 213－224.

[16] AKSNES D W. Characteristics of highly cited papers[J]. Research evaluation, 2003, 12(3): 159－170.

[17] TAHAMTAN I, AFSHAR A S, AHAMDZADEH K. Factors affecting number of citations: a comprehensive review of the literature [J]. Scientometrics, 2016, 107(3): 1195－1225.

[18] LEVITT J M, THELWALL M. Patterns of annual citation of highly cited articles and the prediction of their citation ranking: a comparison across subjects[J]. Scientometrics, 2008, 77(1): 41－60.

[19] WALTERS G D. The citation life cycle of articles published in 13 american psychological association journals: a 25-year longitudinal analysis[J]. Journal of the American Society for Information Science and Technology, 2011, 62(8): 1629－1636.

[20] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3): 851－872.

[21] GLANZEL W, SCHOEPFLIN U. A bibliometric study on aging and reception processes of scientific literature[J]. Journal of information science, 1995, 21(1): 37－53.

[22] ADAMS J. Early citation counts correlate with accumulated impact [J]. Scientometrics, 2005, 63(3): 567－581.

[23] LI J, YE F Y. A probe into the citation patterns of high-quality and high-impact publications[J]. Malaysian journal of library & information science, 2014, 19(2): 17－33.

[24] LI J, YE F Y. Distinguishing sleeping beauties in science[J]. Scientometrics, 2016, 108(2): 821－828.

[25] EGGHE L, RAO I. Citation age data and the obsolescence function - fits and explanations[J]. Information processing & management, 1992, 28(2): 201－217.

[26] MCCAIN K W, TURNER K. Citation context analysis and aging patterns of journal articles in molecular-genetics [J]. Scientometrics, 1989, 17(1/2): 127－163.

作者贡献说明：

熊泽泉：负责数据分析、论文撰写与修改；

段宇锋：负责论文修改与定稿。

Citation Patterns of Chinese Academic Journal Papers: A Case Study of the Journal Articles (2006－2008) in the Field of Library and Information Science

Xiong Zequan[1,2]　Duan Yufeng[1]

[1] Faculty of Economics and Management, East China Normal University, Shanghai 200241

[2] Library, East China Normal University, Shanghai 200241

Abstract: [**Purpose/significance**] This paper explores the citation patterns of Chinese academic journal papers and the influence of time window, and establishes an analysis frame of the citation patterns. [**Method/process**] Taking library and information science for example, a two-step cluster analysis on the absolute citations and the relative citations of Chinese journal papers published between 2006 and 2008 was performed. Correlation between main characteristic factors and citation patterns of papers was analyzed. [**Result/conclusion**] Three patterns were detected from the perspective of absolute citations, and all of them show a classic citation pattern - rise first and then fall. Six citation patterns are detected from the perspective of relative citations, and three of them can be classified as classic citation pattern, while other three clusters can be labeled as quasi-sleeping beauties, positive bias pattern and marathoners. Moderate to high correlations between initial citations and total citations under the relative perspective, and the correlation strengthen with the average citation counts which reversed under the absolute perspective. The results indicate that the correlation between initial citations and total citations depends on the curvature of citation curve but not the number of total citations.

Keywords: citation　citation patterns　time windows　obsolescence　correlation analysis